DOCUMENT RESUME

ED 286 921 TM 870 556

AUTHOR Forster, Fred
TITLE Riding the Rasch Tiger. Part 1: Laying the Item Bank Foundation (Paul Volker Would Approve).
PUB DATE Apr 87
NOTE 38p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, DC, April 21-23, 1987). For parts 2 and 3, see TM 870 557-558.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Difficulty Level; Elementary Education; Item Analysis; *Item Banks; Item Sampling; *Latent Trait Theory; Mathematical Models; *Measurement Objectives; Test Items; Test Reliability; *Test Theory; Test Validity
IDENTIFIERS *Calibration; Point Biserial Correlation; Portland Basic Skills Achievement Levels Testing; *Rasch Model

ABSTRACT

Studies carried out over a 12-year period addressed fundamental questions on the use of Rasch-based item banks. Large field tests administered in grades 3-8 of reading, mathematics, and science items, as well as standardized test results were used to explore the possible effects of many factors on item calibrations. In general, the results confirmed the practical value of Rasch-scaled items. For example, to investigate whether item calibrations were the same for different groups, reading test results from 1,400 fourth graders were calibrated separately for boys and girls, above and below average achievement levels, and inner city versus economically advantaged schools. Item calibrations for each subgroup were strongly correlated (.93-.99) with calibrations for the total group. Two other major empirical issues were the extent of test linkage to a single curriculum scale and the comparability of scores from different tests. Other questions included: (1) whether item bank calibrations can be used to scale new tests; (2) whether item measuring a different latent trait can be detected (by examining point biserial correlations and item curves); and (3) whether item calibrations are appreciably affected by test range, item discrimination, sample size, grade level, and time of year. (LPG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED286921

RIDING THE RASCH TIGER

PART 1: LAYING THE ITEM BANK FOUNDATION
(Paul Volker Would Approve)

Fred Forster

Microsystems for Education and Business, Inc.
395 S.W. 6th Street
Gresham, Oregon 97080
(503) 666-7883

Presented at
The National Association of Test Directors Symposium
The National Council on Measurement in Education
Annual Conference
April 1987
Washington, D.C.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

F. Foster

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 870 556

BEST COPY AVAILABLE

RIDING THE RASCH TIGER

PART 1: LAYING THE ITEM BANK FOUNDATION
(Paul Volker would approve.)

Introduction

The 1986 NATD Symposia included Taming the Rasch Tiger which raised several important issues on the use of the Rasch model in the public schools. This year's theme, Riding the Rasch Tiger, is intended to help you harness the Rasch model to develop better tests for your district.

The Rasch model makes item banking practical and sensible. With all other measurement models, what we know about an item is mathematically tied to what we know about the group which took it. With the Rasch model, we have the potential to establish curriculum scales in each subject which standardize our knowledge about an item. At least, that is what the Rasch experts say. (The theoretical basis of the Rasch model is discussed in Wright, 1968, Wright & Panchapakesan, 1969, and Wright and Stone, 1979; a school oriented view of this same topic appears in Ingebo, 1980; a discussion of the role of the Rasch model in item bank development appears in Forbes, 1986, and Hunt, 1986.)

How do we know that what the experts say is true? What issues must be addressed to show that the Rasch model can be used to lay the foundation for item banking? (Historical summaries of the development of item banking by Portland (Oregon) Public Schools (P.P.S.) and the Northwest Evaluation Association (N.W.E.A.) appear in Ingebo, 1978 and Hathaway, Houser, & Kingsbury, 1985.)

Are Item Calibrations the Same for Different Groups?

This is the fundamental question of the Rasch model. A potpourri of testing techniques is available to examine items and analyze test data. Only the Rasch model claims to produce calibrations that are mathematically independent of the groups that were tested. If the answer to this question is "yes," then the Rasch model can help in item bank development. If the answer is "no," then it is necessary to look for help elsewhere.

Since we are from Oregon (not all that far from Missouri), we wanted to answer this question with our own student scores (Forster, 1978a). Our first study compared calibrations for 1,400 fourth grade students on a standardized reading test. We calibrated the items for boys and girls scoring above and below the average. Then we calibrated the items for students attending inner city and "advantaged" schools. The results are shown in Table 1. The scatter-plots were tight and the lowest correlation of the calibrations between groups was 0.93.

If you are familiar with the literature on the Rasch model, you probably have encountered the item calibrations expressed as log units (logits), positive and negative values centered on 0. In practice, we use Rasch units (Rits) which center on 200 and range roughly from 130 to 270 for a given subject such as reading or mathematics. (One

BEST COPY AVAILABLE

Rit equals .1 logit.) Rits are convenient because they are always positive integers and are not easily confused with percentiles or standard scores.

We replicated this study using scores on standardized tests for 4,000 students in each of grades four and eight. For this study we did not calibrate boys and girls separately, but we did divide the population into groups scoring either above or below the average and into groups representing inner city schools and higher scoring schools. As shown in Table 2, the correlations of the calibrations in this study were slightly higher than those in the first study. This was probably the result of the increased accuracy of the calibrations due to the larger sample sizes.

Recently, this study was repeated for San Francisco Public Schools for a ninth grade competency test developed by the Southwest Regional Educational Laboratory. The population was divided into those scoring below the average and those scoring above the average. As shown in Table 3, the change of scenery did not seem to change the results.

Bruce Arneklev (1979) of Tacoma Public Schools repeated this research adding a twist. He separated students into different ethnic groups and then correlated the calibrations across groups. As you would expect, he found significant differences in the averages and distributions, but the correlations between the calibrations were comparable to those reported in Tables 1 to 3.

If you are superstitious, these results look a little too good to be true, and maybe they are. While the plots of the calibrations for different groups are tight, there is drift at the bottom of the plots for high groups and at the top for low groups, i.e., where the tests are least appropriate for the group. (Arneklev also found drift for ethnic groups where you might suspect a language problem.) So there is the hint that the Rasch is human enough to have some limitations. These data suggest that we should look for the possibility of calibration drift (inaccuracy) for items that are much too easy or much too hard, that cover material which has not been taught, or that contain language which is unfamiliar to students. (The phenomenon of calibration "drift" is discussed in Thomas, 1986, and Houser & Kingsbury, 1986.)

Do Tests Link to a Curriculum Scale?

This question is essential to using the Rasch model for item banking. According to the experts, the Rasch model is supposed to calibrate items in relation to the curriculum for a subject such as reading or mathematics. (Forbes, 1986, discusses this relation of calibration to curriculum.) If these calibrations are consistent, then the Rasch has proven its worth as an item banking tool. If the calibrations are inconsistent, the Rasch model will have limited utility for item banking. (For a discussion of the process of "linking" tests to compare calibrations to a fixed scale see Forster, 1978b; for a discussion of the effect of the restriction of item range on linking see Ingebo, Arneklev, & Gee, 1976.)

It takes some planning to answer this question. Several tests must be constructed with overlapping items to test whether the item calibrations match up. The trick is to design enough overlap in the right places so the tests can be "linked" to a single scale. Fortunately, George Ingebo and Dean Forbes (1975) of P.P.S. took on this challenge.

Ingebo developed an ingenious plan for the research study using 250 questions that had been field tested with seventh graders. First he separated the questions with p-values under 20% (hard) from those with p-values over 20% (easier). The easier items were formatted into four 60-item tests (W, X, Y, and Z) with a 20-item overlap between

Table 1

Item Calibrations for Different Sex and Achievement Groups -- Fourth Grade Reading

Comparison	Correlation
Girls above the average vs. Total	.98
Girls below the average vs. Total	.96
Boys above the average vs. Total	.97
Boys below the average vs. Total	.95
Inner City Schools vs. Total	.99
Advantaged Schools vs. Total	.93
Inner City Schools vs. Advantaged Schools	.97

Table 2

Item Calibrations for Different Achievement Groups

Group	4th Reading	4th Math	8th Reading	8th Math
Above the average vs. Total	.99	.98	.98	.97
Below the average vs. Total	.96	.97	.98	.95
Inner City vs. Total	1.00	.99	1.00	1.00
Advantaged vs. Total	1.00	.99	1.00	1.00
Inner City vs. Advantaged	.99	.98	.99	.98

Table 3

Item Calibrations For Different Achievement Groups on the San Francisco Reading Competency Examination

Group	Correlation	Average Absolute Value Difference
Above the Average vs. Total	.96	2.43
Below the Average vs. Total	.99	0.90
Above the Average vs. Below the Average	.92	3.19

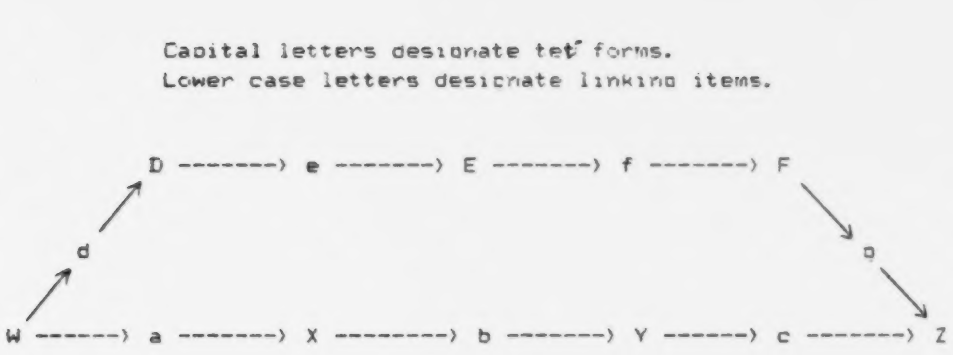

Figure 1. Linking network for seventh grade field testing.

Table 4

Linking Values for the Iroquois Seventh Grade Mathematics Test Network

First Form	Second Form	Linking Items	Average First Form	Average Second Form	Linking Value
W	X	a	207.5	195.4	-12.1
X	Y	b	204.5	197.6	- 6.9
Y	Z	c	202.5	196.1	- 6.4
W	D	d	207.0	186.9	-20.1
D	E	e	206.9	203.7	- 3.2
E	F	f	203.4	200.4	- 3.0
F	Z	g	194.0	196.0	2.0
LINKING VALUE W--)X--)Y--)Z (a+b+c)					-25.4
LINKING VALUE W--)D--)E--)F--)Z (d+e+f+g)					-24.3

Table 5

Rit Score Averages for Students on Short Field Tests and Longer Standardized Tests

Grade	Field Test Form	Average Raw Score		Average Rit Achievement Scores	
		Field Test	Standardized Test	Field Test	Standardized Test
3	R12	18.1	16.2	186.8	187.2
4	R17	19.6	33.5	194.9	194.6
4	R35	19.6	36.1	195.8	196.8
5	R36	15.6	31.7	199.3	199.1
6	R25	14.4	38.6	204.4	205.3
6	R37	11.0	38.4	206.2	205.3
7	R27	14.1	41.0	213.1	211.4
7	R38	12.9	35.7	206.6	207.4
8	R32	12.7	61.8	211.7	211.7
8	R39	15.0	57.5	211.2	209.4

Table 6

Comparison of Test Scores Between the Competency Progress Achievement Level Programs

Reading			Mathematics		
Grade	Competency Test Average	Survey Test Average	Grade	Competency Test Average	Survey Test Average
4	194.12	193.76	4	191.28	190.36
5	200.00	200.00	5	200.81	200.00
6	205.53	205.74	6	207.22	208.36
7	209.77	211.21	7	213.49	214.87
8	214.74	216.83	8	222.15	221.41

adjacent tests. The hard items were formatted into three 60-item tests (D, E, and F) with a 20-item overlap between adjacent tests. Then the pairs of tests at the extremes (W and D, and Z and F) were given a 20-item overlap (see Figure 1). This linking pattern provides two independent ways to link tests W and Z. The first is W to X, X to Y, and Y to Z. The second is W to D, D to E, E to F, and F to Z.

Ingebo and Forbes planned the administration so seventh graders took the easier tests and eighth graders took the harder tests. The calibrations for the items in the overlap between tests were "linked," i.e., plotted and averaged. The linking value, i.e., the amount to be added to the calibrations on the first test to bring them in line with the second test (see Table 4), was calculated as the difference in the average for the calibrations of the items shared by the two tests. Adding the linking values W to X (-12.1), X to Y (-6.9), and Y to Z (-6.4) produced -25.4 Rits. Adding the linking values W to D (-20.1), D to E (-3.2), E to F (-3.0) and F to Z (2.0) produced -24.3 Rits. This represented a 4.4% difference in the independent linking values. (Checking the accuracy of linking using this "triangulation" method is discussed in Karr, 1979.)

This landmark study showed that linking works when things are done just right. Later, we found there were times when linking wasn't nearly as smooth or accurate. Our worst horror story came from the analysis of a group of five field tests given in several Oregon districts by the Oregon Department of Education. Four tests designed for the fourth grade curriculum had an overlapping pattern similar to the original study by Ingebo and Forbes. When the two linking values were calculated, they differed by 15 Rits, about three times what we considered acceptable. After checking the screening reports, we found the problem. Two of the schools "helped" by giving the tests to their top eighth grade students. In those schools, the students "cooled" the test and the p-values for the items ranged from 93% to 98% which raised havoc with the calibrations. (The problem of estimating calibrations for extreme p-value items is discussed in Forster, 1978f.) When these spurious scores were dropped from the analysis, the linking values fell back into line and the actual values were confirmed.

Linking has proven to be sensitive to problems in test administration, curriculum, and language. As a result, George Ingebo (1976) developed the concept of constructing field tests to match a "linking network" pattern. This pattern specifies the overlap of items to ensure that each question is calibrated on at least three different tests and each link between tests is checked at least three different ways. The development of linking network patterns made it possible for P.P.S. and N.W.E.A. to coordinate large scale field testing of thousands of items on a national basis and still maintain tight quality control. Even this technology doesn't guarantee that everything will go right; it only guarantees that you know when things start to go wrong.

Experience has shown there is no simple statistical way to handle discrepancies in linking. Encouraging a computer to apply a multivariate-whatever analysis blindly on data that do not agree averages error throughout your item bank. Linking tests without at least three cross-checks to uncover potential discrepancies is dangerous. There is no substitute for the human touch of eye-balling each printout, graph, and linking report to identify and eliminate error before it gets a foothold. Even then, you may need every bit of your creativity to discover the cause and cure for problems that arise. As George Ingebo (personal communication, 1980) once pointed out, reading the first 10,000 or so printouts is the hardest. (A more efficient method for calibrating using existing item banks is discussed in Forster, 1982a; Forster, 1982b; and Forster & Karr, 1986.)

Do Students Score the Same on Different Tests?

This question is more complicated than it may seem. We should find students have comparable scores for tests that are comparable. But there are many situations where students should be expected to score differently on different tests. For example, several researchers have reported attempts to "vertically equate" tests. Generally in these studies, students take two or more wide range standardized tests designed for different grade levels, and the data are used to build an "equating table" between the test forms. And this creates a problem.

Is it reasonable to expect a student to score comparably on tests when one is appropriate and the other isn't? In the worst case, both tests may be inappropriate for the student. In the best case, the student will take at least one inappropriate test. With this in mind, it isn't surprising to find that the results of vertical equating have been inconsistent. The mis-match between the students' achievement level and the two tests is probably a major contributor to the inconsistencies reported by Holmes (1982), Rentz & Bashaw (1977), and Slind & Linn (1977).

What is reasonable to expect? If the experts are correct, we should be able to have students take two different appropriate tests and the scores should differ by no more than the standard error of measurement. To check this out you need to compare results for two tests already calibrated to the same scale. (See Forster, 1978c.)

Fortunately, we had the results from several short, 30-item reading field tests which had been given to students who had previously taken standardized reading tests of 80 to 100 items. Both sets of tests were linked to the P.P.S./N.W.E.A. reading item bank scale so the student Rit scores could be compared. The results are shown in Table 5.

As these data show, the average Rit scores were quite similar between the different tests. We found, however, that the "ceiling" and "floor" of the short tests affected these averages. When a large number of students scored at the extremes, then the average tended to be a Rit or two above or below that for the standardized test.

This study was repeated using two series of 40-item tests. One series provided competency tests designed for each grade three to eight. The other series assigned each student a test matched to his/her achievement level. The results are shown in Table 6. As with the first study, the averages agreed except where the competency tests created a floor or ceiling effect.

A similar study was performed for San Francisco Public Schools. This study compared scores for ninth grade students on the San Francisco competency test to their scores on the P.P.S. reading competency test. Although the tests were comparable, the Portland test spans a greater achievement range, while the San Francisco test has an effective ceiling around a Rit score of 230. The Rit averages for the group were 214.51 on the San Francisco test and 214.27 on the Portland test, which compare favorably. As the scatter-plot in Figure 2 shows, however, the agreement decreased as the scores approached the ceiling for the San Francisco test. As a result, the average absolute value difference in the scores was 7.26, a level far from optimum. These data indicate that the comparability of scores was influenced by the match between each test and the group being tested.

Another study done for Hillsboro (Oregon) Public Schools compared scores from level tests in the fall to an end of first semester test in general mathematics for 303 eighth grade students. The scatter-plot of these data, shown in Figure 3, shows that the

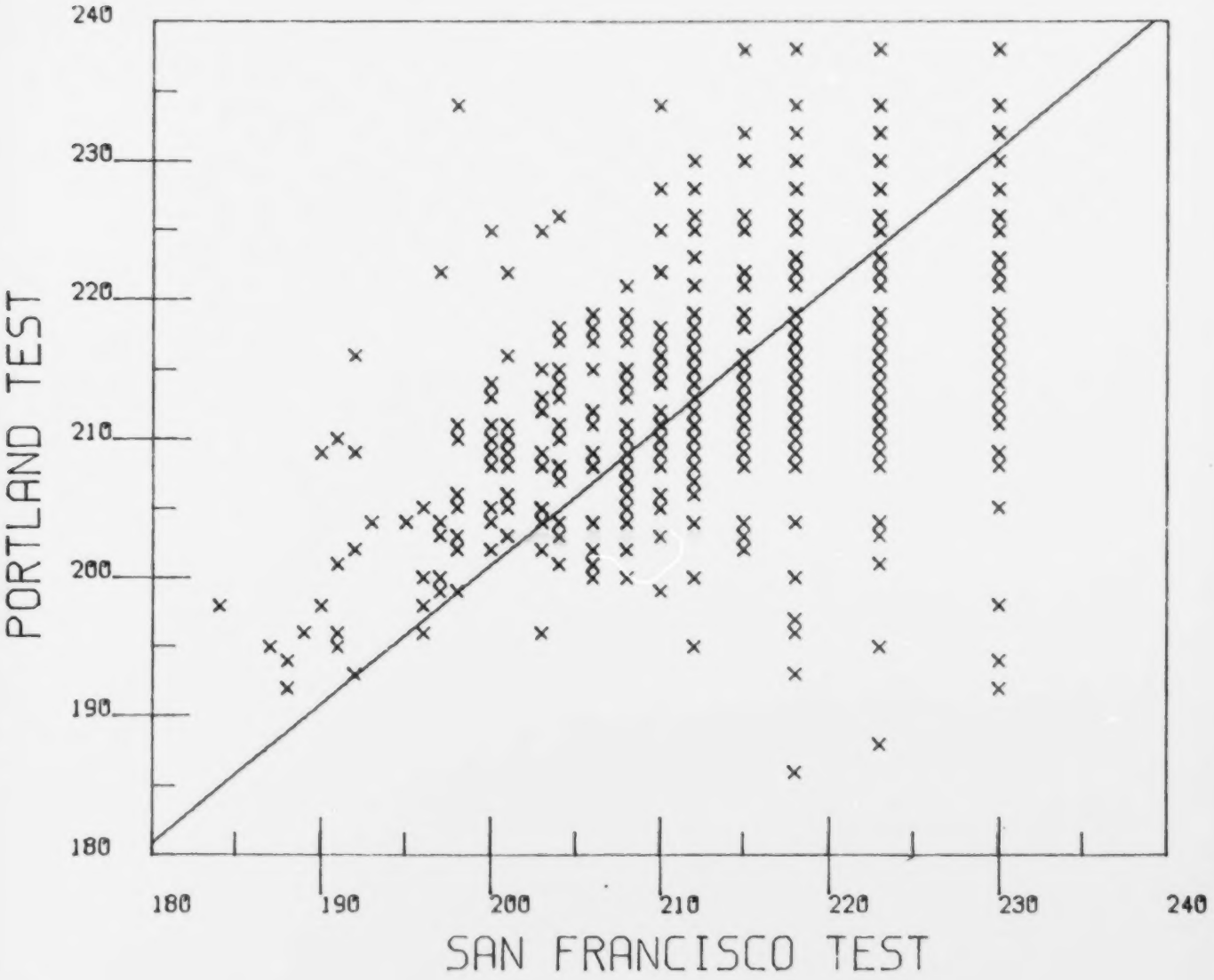

Figure 2. Comparison of RIT scores. San Francisco and Portland Competency Tests.

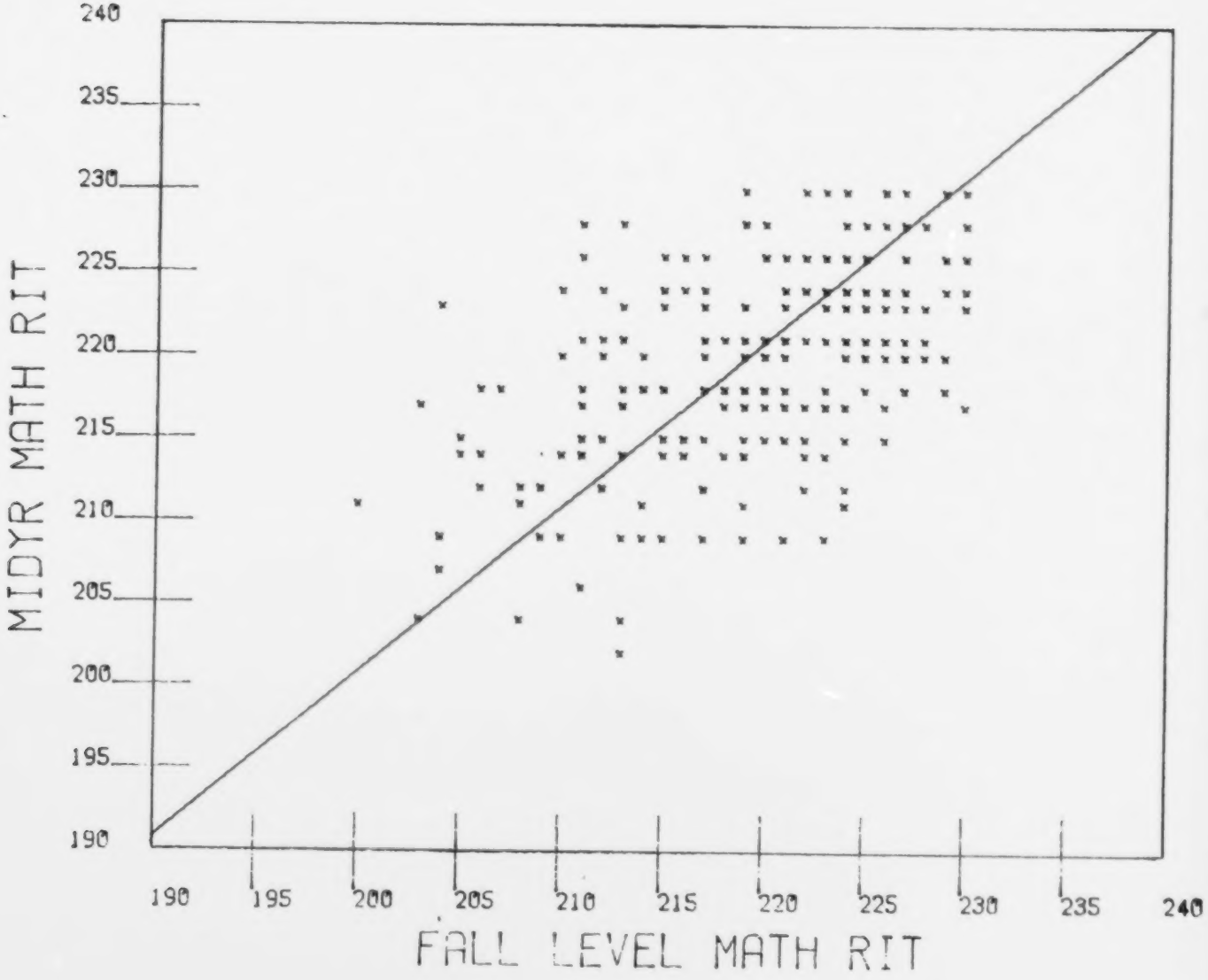

Figure 3. Comparison of RIT scores. Fall level tests and 8th grade midyear general mathematics.

test scores were similar but not identical. (In this instance, the average absolute value difference was 4.7 Rits, significantly less than observed in the San Francisco study.)

As these studies show, the comparability of scores depends on the comparability of tests and the match to students. When students top out on one of the tests, their scores are depressed. When students bottom out, their scores are inflated. Although the Rasch averages for groups may be comparable, these results point out the importance of choosing tests well matched to a student's achievement level to ensure comparability of scores. (For a detailed discussion of the techniques for matching tests to students see Forster, 1975a; Forster, 1986; and Hathaway, et al., 1985.)

Other Questions

The positive answers to the crucial questions we have already discussed establish the usefulness of the Rasch model for item banking. The questions we will discuss now add the frosting to the cake.

Beginners in the item banking game are fond of talking about the "assumptions" of the Rasch model. It is important not to confuse these discussions with the empirical question of whether the model works. For example, no one expects an actual gas to fit the assumptions of the perfect gas law exactly. Scientists and engineers gauge the value and utility of this mathematical model by its ability to predict pressure, volume, and temperature in real world situations. Similarly, the proof of the Rasch model (or any measurement model) lies in its ability to produce calibrations, scores and parameters that are consistent across time and testing situations.

As no gas actually fits the perfect gas law, it is a safe bet that no item exactly fits the Rasch model. In fact, it seems unlikely that any reasonable mathematical equation will ever capture the ingenuity and motivation of youngsters except as averaged across large groups. So the statistical controversies over hypothetical model fit are little more than arguments about a tautology. And these arguments are no substitute for the scientific examination of real issues such as consistency of calibration, test scaling, and the accuracy of student scores.

A related issue is the substitution of computer simulated data for test data actually drawn from a school testing program. It is remarkable that reviewers don't reject these studies out of hand. Mathematical concoction is, at best, an inadequate reflection of partial and incomplete knowledge of the performance of real students. If we hope our research on mental measurement to be respected by the broader scientific community, we must repudiate these non-scientific approaches and address the issues with real data.

And now, what other questions will help us ride the Rasch tiger as we build item banks?

Can Rasch Scores be Predicted?

This question is a logical consequence of the notion of an item bank curriculum scale. To be really flexible, item calibrations should provide the information necessary to scale new tests that are developed from an item bank. If the answer to this question is

yes, we can use item bank calibrations to scale new tests. If the answer is no, it is necessary to try out the new tests before using them. (See Forster, 1976.)

We devised a study to answer this question using follow-up data to the Ingebo and Forbes (1975) linking study reported above. Using the linking values and items from that study, a small item bank was built. Then seven new 25-item mathematics tests were built from the bank. Each new test overlapped 12 items with the next lower test, 13 items with the next higher test. The new tests were given to a new group of 1,500 seventh grade students.

The new tests were scaled using the Rasch model and the field test information. The tests were also scaled from the item bank. The correspondence between these independent scalings for one of the tests is shown in Table 7. (The results for all seven tests were similar.) As these data indicate, the difference between the two scalings was small in comparison to the scale and the standard error of measurement.

An interesting variation of this study was reported by Clarence Mershon (1986). He found that the Portland Achievement Level Test (P.A.L.T.) program made it possible to scale the Parkrose goal-based tests and provide precise information about the sources of the inconsistencies which had been noted by teachers. This also made it possible to engineer the Parkrose tests to eliminate these problems.

What is the Latent Trait?

The term latent trait is an unfortunate verbal accident that has stuck. It was originally intended to indicate that the variable we are measuring is not directly observable, but must be inferred. Yet we don't use the term latent when considering variables such as temperature or barometric pressure which are inferred from our measuring instruments just as mathematics or reading achievement are inferred from tests. In fact our belief in a latent trait or underlying curriculum variable is implicite whenever we test. For example, what sense would it make to give a mathematics test unless we thought that a higher score on the test was an indication of higher math achievement.

There is an intriguing practical question associated with the concept of a latent trait. Can the Rasch model be used to tell if some of the questions on a test belong to a different latent trait? Can the Rasch model pick out questions which are measuring something else? (A discussion of the effect of the latent trait on linking appears in Forbes & Ingebo, 1975.)

We devised a study to use the Rasch model to unscramble items purposely mixed between two different subject areas. We used science and mathematics test scores for 450 students in the Hillsboro District in the seventh grade in the spring of 1986. We matched a 47-item science test to a 40-item general mathematics test. Then we analyzed 10 of the science items and 10 of the mathematics items in relation to: (1) the full science test, (2) the full mathematics test, (3) a contrived test of 40 mathematics items and the 10 science items, (4) a contrived test of the 10 mathematics items and 40 of the science items, and (5) a contrived test of 40 mathematics items and 40 science items. Would the Rasch model pick out the questions which didn't belong?

This study required a method for separating the good items from the bad and the ugly. One tried and true technique is the point biserial. If an item has a low point biserial (but the p-value is okay) then this may indicate a lack of fit to the latent trait.

Table 7

Comparison of Scaling from Item Bank Calibrations and Field Test Calibrations

Raw Score	Predicted Scaling	Actual Scaling	Difference in Scaling
1	167.8	167.2	.6
2	175.2	174.7	.5
3	179.7	179.3	.4
4	183.1	182.7	.4
5	185.9	185.5	.4
6	188.2	187.9	.3
7	190.3	190.1	.2
8	192.3	192.1	.2
9	194.1	194.0	.1
10	195.8	195.8	.0
11	197.5	197.5	.0
12	199.2	199.2	.0
13	200.8	200.9	.1
14	202.5	202.6	.1
15	204.2	204.3	.1
16	205.9	206.1	.2
17	207.7	207.9	.2
18	209.7	209.9	.2
19	211.8	212.1	.3
20	214.1	214.5	.4
21	216.9	217.3	.4
22	220.3	220.7	.4
23	224.8	225.2	.5
24	232.1	232.6	.5

Table 8

Identification of "Mixed" Items

	SCIENCE ITEMS		
	Full Science Test	40 Mathematics and 10 Science Items	40 Mathematics and 40 Science Items
Item	Point Biserial	Point Biserial	Point Biserial
1	0.32	0.17	0.22
2	0.30	0.19	0.22
3	0.29	0.18	0.20
4	0.21	0.15	0.14
5	0.37	0.22	0.23
6	0.37	0.19	0.27
7	0.36	0.15	0.19
8	0.33	0.22	0.25
9	0.39	0.25	0.32
10	0.36	0.21	0.27

	MATHEMATICS ITEMS		
	Full Mathematics Test	40 Science and 10 Math Items	40 Science and 40 Mathematics Items
Item	Point Biserial	Point Biserial	Point Biserial
1	0.26	0.21	0.25
2	0.14	0.13	0.14
3	0.46	0.31	0.39
4	0.47	0.42	0.45
5	0.45	0.35	0.40
6	0.31	0.25	0.28
7	0.25	0.22	0.23
8	0.28	0.28	0.28
9	0.32	0.33	0.32
10	0.46	0.32	0.40

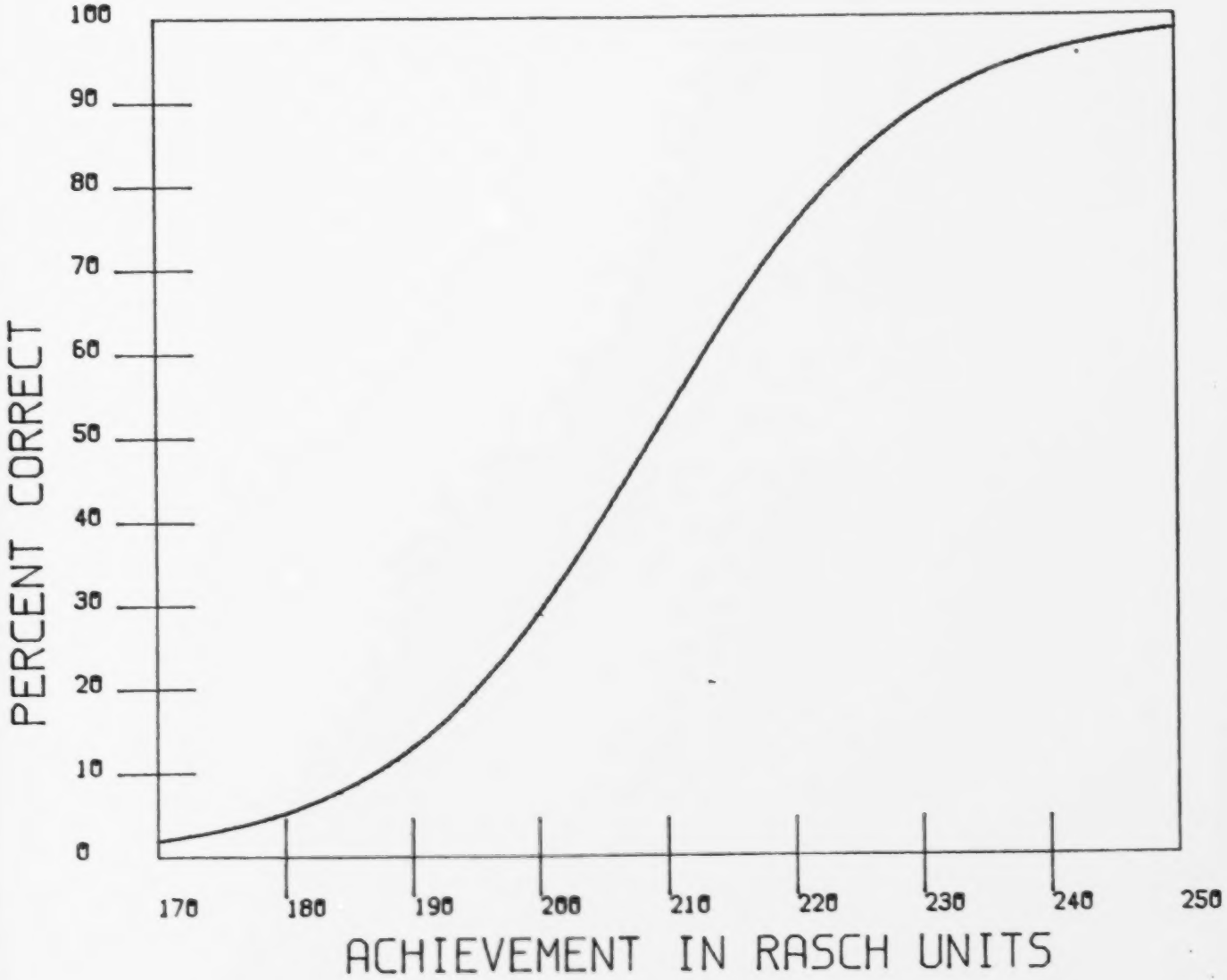

Figure 4. Theoretical Rasch item curve. San Francisco competency exam.

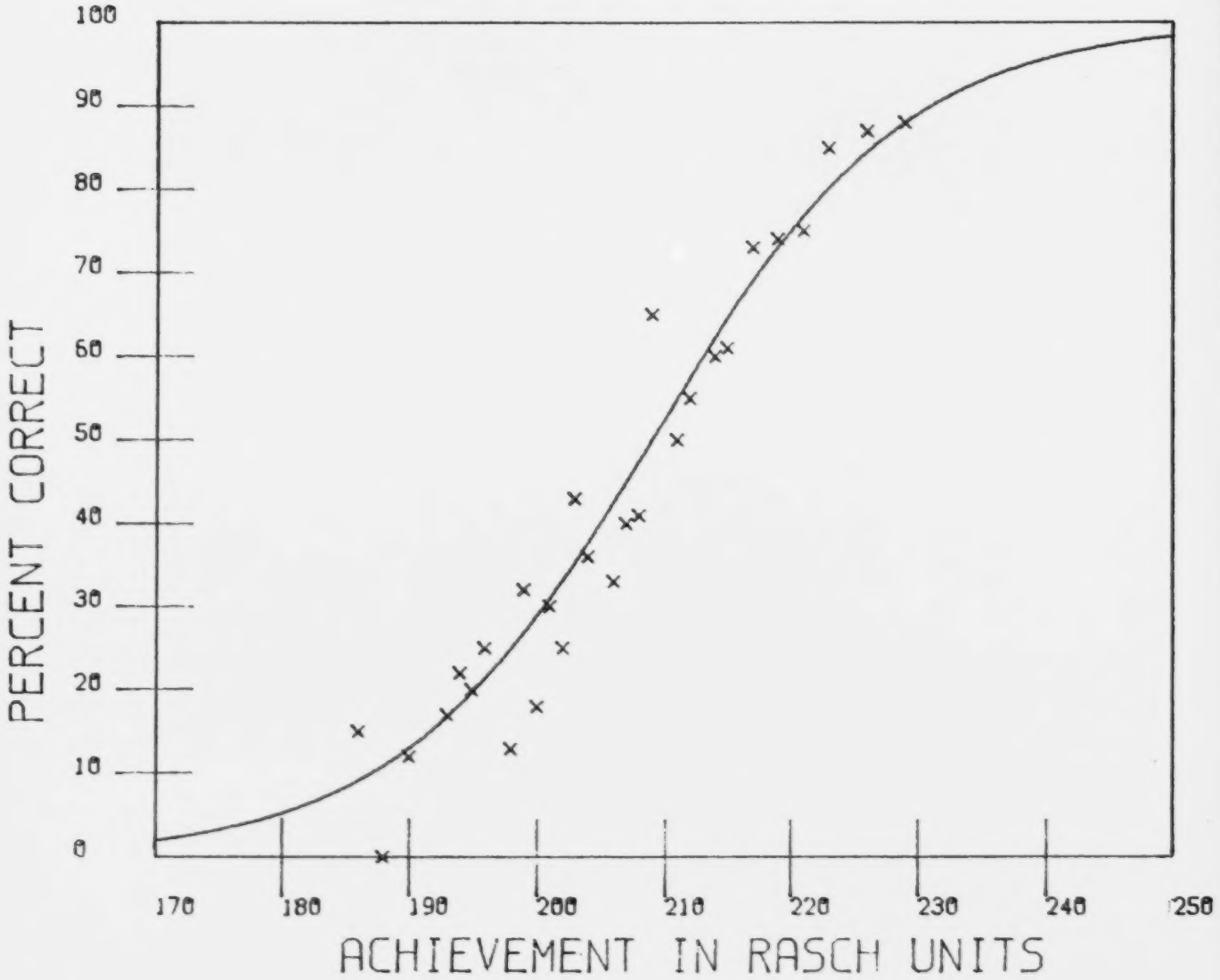

Figure 5. Actual Rasch item curve. San Francisco competency exam.

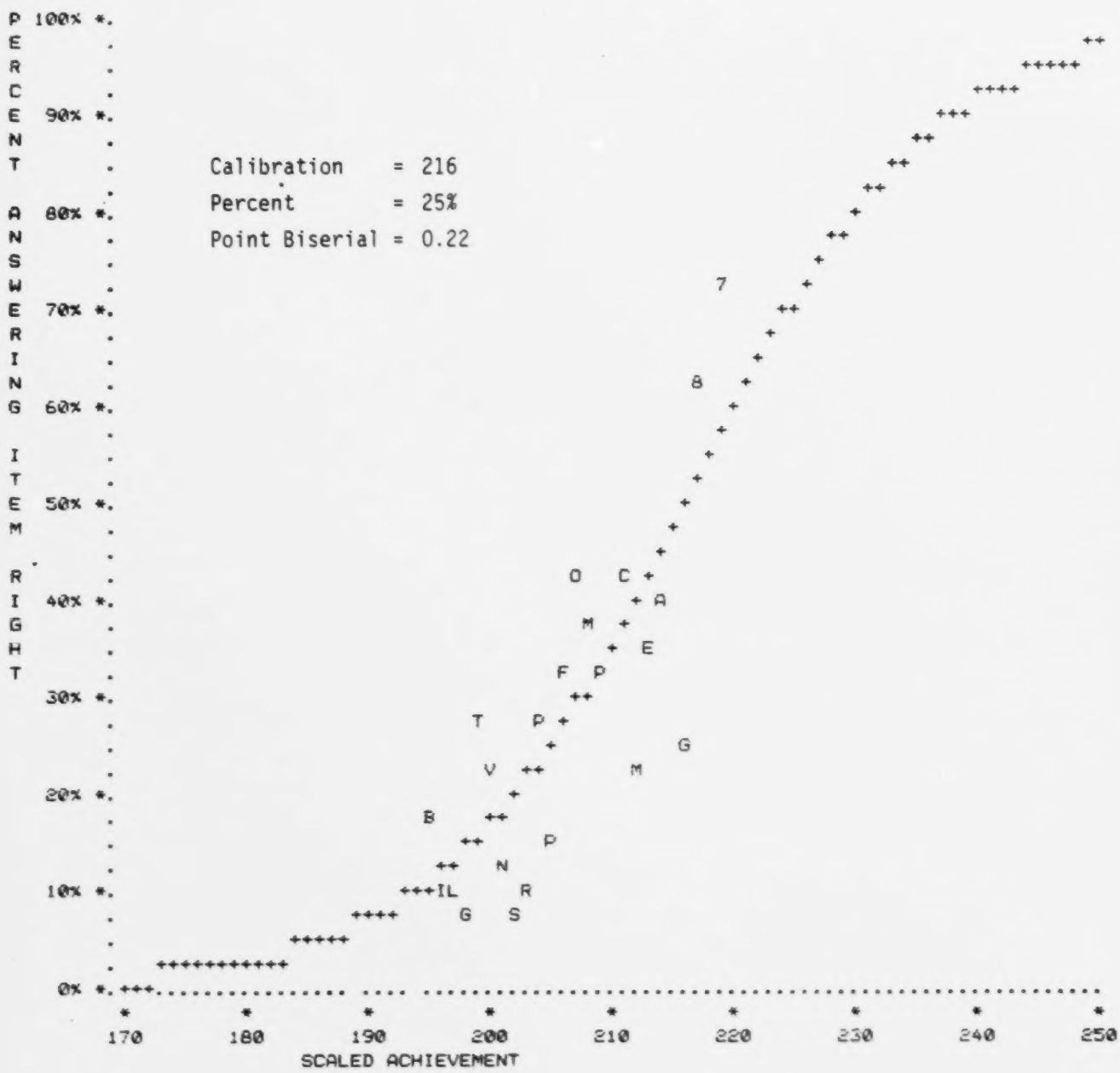

Figure 6. Example of an item with a low point biserial and acceptable item curve.

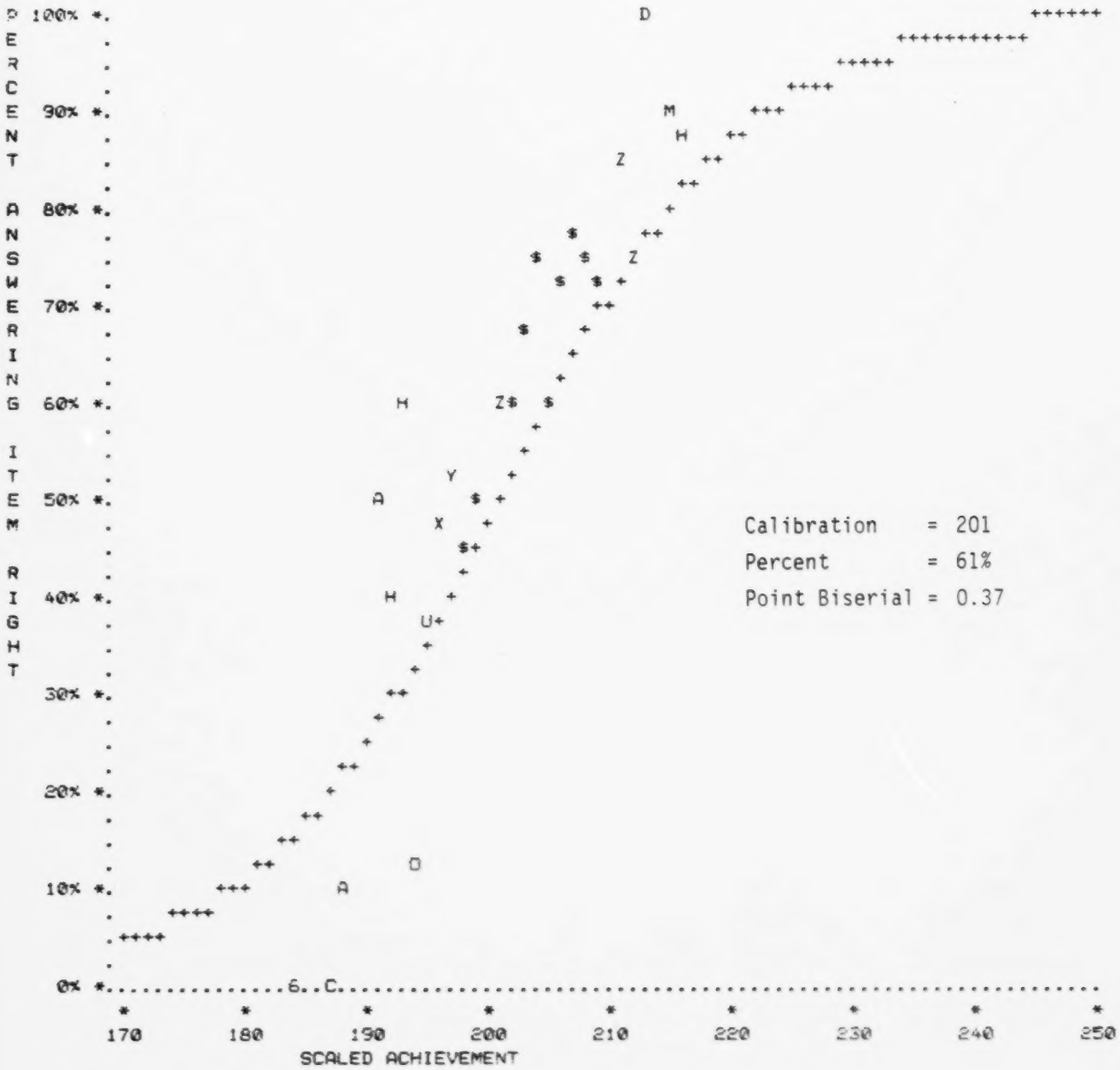

Figure 7. Curve for example science item on full science test.

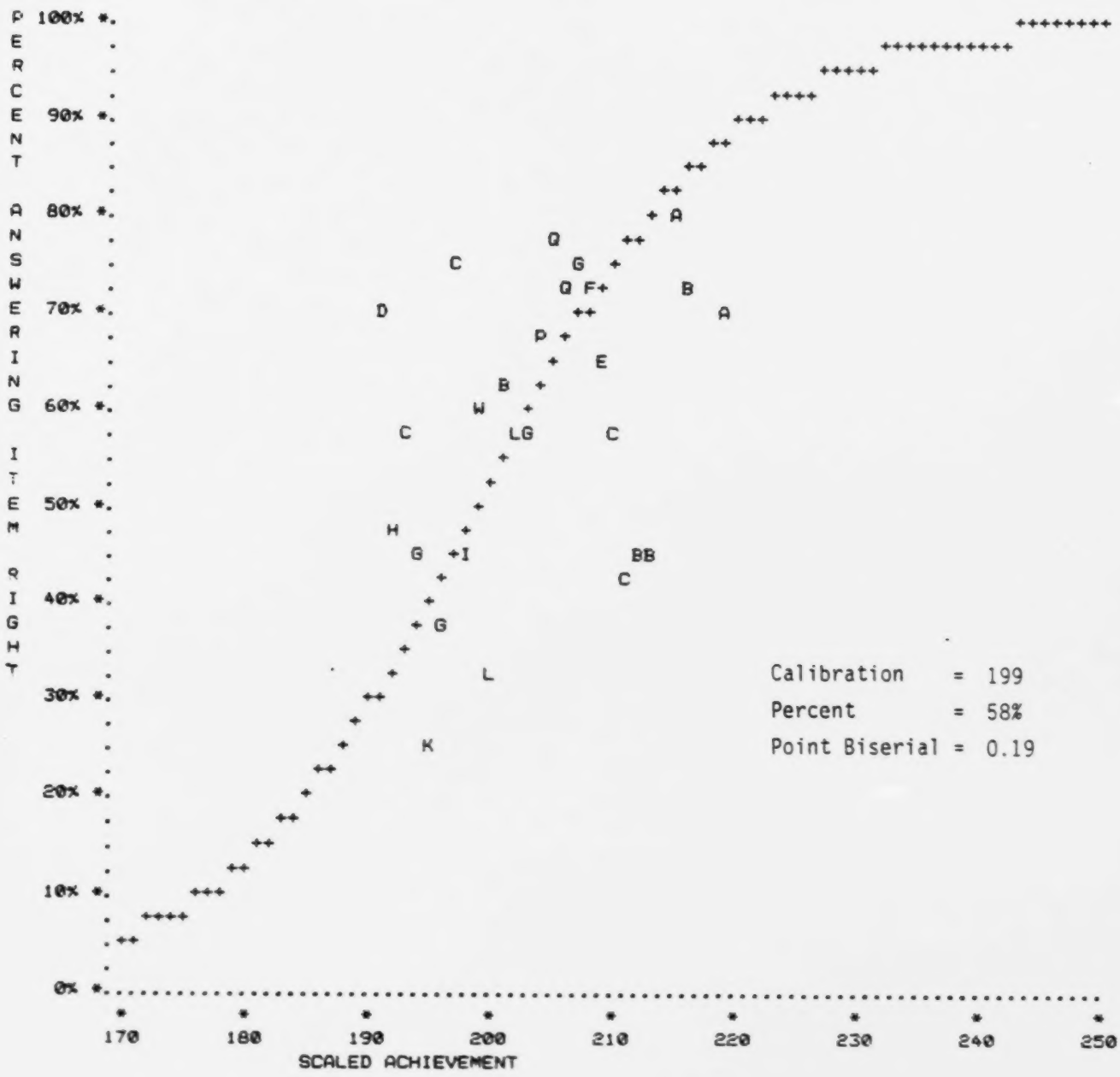

Figure 8. Curve for example science item on "test" with 40 mathematics and 10 science items.

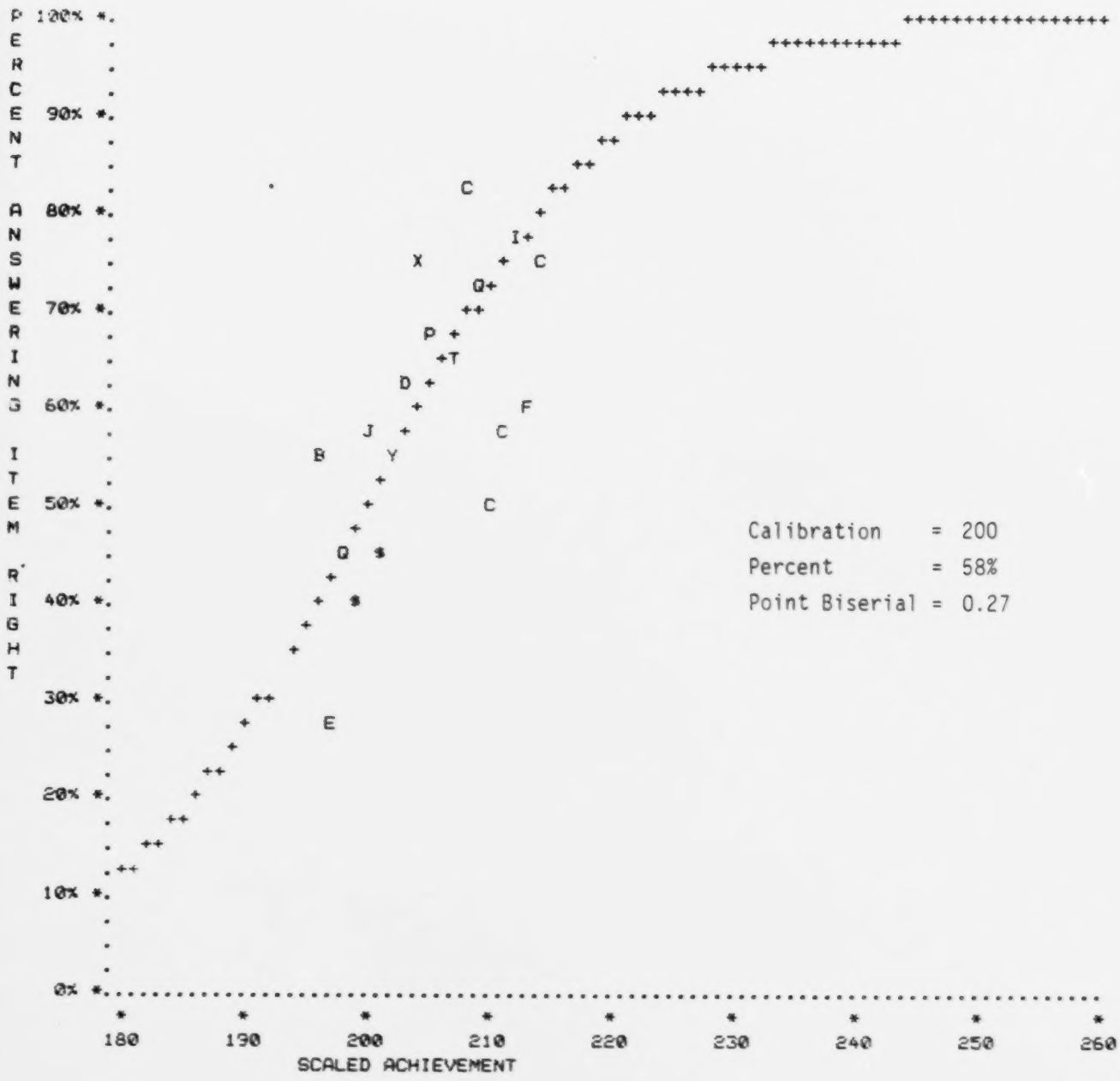

Figure 9. Curve for example science item on "test" with 40 mathematics items and 40 science items.

The point biserials were calculated for the 10 science items and the 10 mathematics items under the three conditions: (1) in relation to the "pure" test, (2) in relation to themselves and 40 items from the other subject, and (3) in relation to themselves and 30 items from the same subject together with 40 items from the other subject. As shown in Table 8, the point biserials dropped as more items from the other subject were included. And this effect was significantly greater for science than for mathematics.

You may wonder why we used the point biserial as a part of this study. The answer is simple; it provides valuable information about an item. (Real test developers do eat quiche and use the best available proven statistics whether Rasch or "classical.")

The Rasch model provides a tool, the item curve, which is ideal for this job. Using the Rasch equation, we can make a theoretical estimate of what proportion of students, at each achievement level, should get the question right. This leads to the theoretical item curve with its characteristic "S" shape as shown in Figure 4. Then we can check the performance of students, at each level, to superimpose the actual item curve shown in Figure 5. If the actual data follow the theoretical curve, we feel good about the item. If the actual data don't follow the theoretical curve, then the item is called into question. All we need are enough data to check the item at each achievement level.

What's so great about the Rasch item curve? Well, it is relatively sample independent. For example, the point biserial works fine if there is a large variance of scores, and the p-value of the item is close to 50%. Otherwise, it can indicate an item is bad when it would work fine with a different group.

In contrast to the point biserial, the Rasch item curve provides a working picture for every question, regardless of the variance or p-value. As shown in Figure 6, an item can have a point biserial below 0.30 (a commonly used cutoff) and still fit our expectations for a properly working item. There is still some relativity in judging the item curve. If the items are much too easy or much too hard for the group (such as the 25% item in Figure 6), there is less information to judge its performance than when the item fits "just right."

We examined the item curves for the 10 science questions and the 10 mathematics questions. Figures 7, 8 and 9 show the effect observed for a typical question. The curve in Figure 7 based on the full test is compact and shows some steepness. The curve in Figure 8, shows a dispersion (snowball effect) when the item is related to mostly mathematics items. Finally, the curve shown in Figure 8 shows less dispersion where the number of mathematics and science items are equal. (The letters and numbers in the plot indicate the number of students at that Rit level A=10, B=11, ..., Z=35 and $=36 or more. The larger this group size, the more stable the data point.)

While these data illustrate how Rasch methodology can help identify items that do not fit the theoretical curve, they are not sufficient to indicate whether this is due to a subject matter problem or the item is just poorly constructed. That judgment requires the examination by expert curriculum specialists and item writers.

These data illustrate the agreement we usually find between the results of applying the Rasch model and the classical model. For example, the point biserial and Rasch item curve provide complementary information to evaluate an item's measurement quality.

Some statisticians feel the item curve is messy. They prefer a statistical formula which adds up the differences between the actual and theoretical, the mean square fit. After years of investigation, we have dropped the mean square because it weights small differences at the top and bottom of the item curve too much, it increases as sample sizes increase regardless of the fit of the item, and it is a test of misfit rather than fit. As a test of misfit, it is a tautology (no item fits the model exactly). By increasing the sample size, you can generate a large enough mean square to reject virtually any item.

The mean square fit was the basis for a recent article by Divgi (1986) with the rather immodest conclusion that "The Rasch model doesn't work with multiple choice questions." This author made the mistake of focusing on the mean square statistic instead of analyzing the data to determine the consistency of the scaling and calibration.

Does Test Range Affect Calibration?

The range of item levels on a test should match its purpose. The familiar wide-range publisher's survey test is designed to fit all students in a given grade or in a few grades. A more efficiently focused test can use a more narrow range of items so nearly all are appropriate for a student. Will the Rasch model calibrate the items differently because of restriction in range? (See Forster, 1978e; Forbes & Ingebo, 1975.)

The study we devised used 1,400 scores on an 80-item fourth grade reading test. First, the total test was calibrated. Then the 10 hardest and 10 easiest were dropped to make a 60-item test. Then this procedure was repeated to produce 50, 40, 30, 20 and 10-item tests of decreasing range. The calibrations were correlated requiring equal variances. The results shown in Table 9 indicate the effect of restricting the range was negligible.

Does Item Discrimination Affect Calibration?

Item discrimination is an intriguing topic. Why do some items produce sharp differences between groups at different levels and others produce only gradual differences? Does this property of an item hold up for different groups and different tests? What is the curriculum significance of discrimination? Are there learnings that are go/no-go and others that evolve slowly?

On a practical level, we need to know how discrimination might affect calibration. The Rasch model ignores differences in item discrimination. Does this limitation have an effect on the item calibration?

To study this question we used 500 test scores for eighth grade students on an end-of-year general mathematics test. First, we calibrated all items. Then we calibrated the items with point biserials below the median point biserial. Then we calibrated the items with point biserials above the median. Finally, we plotted the calibrations which are shown in Figures 10 and 11. The correlations were 1.00 for both plots, but there was a slight constriction in the range of the calibrations for the items with the lower discrimination (the standard deviation shifted from 14.1 to 13.7), and a slight expansion in the range of calibrations for the items with high discrimination (the standard deviation shifted from 5.6 to 6.0).

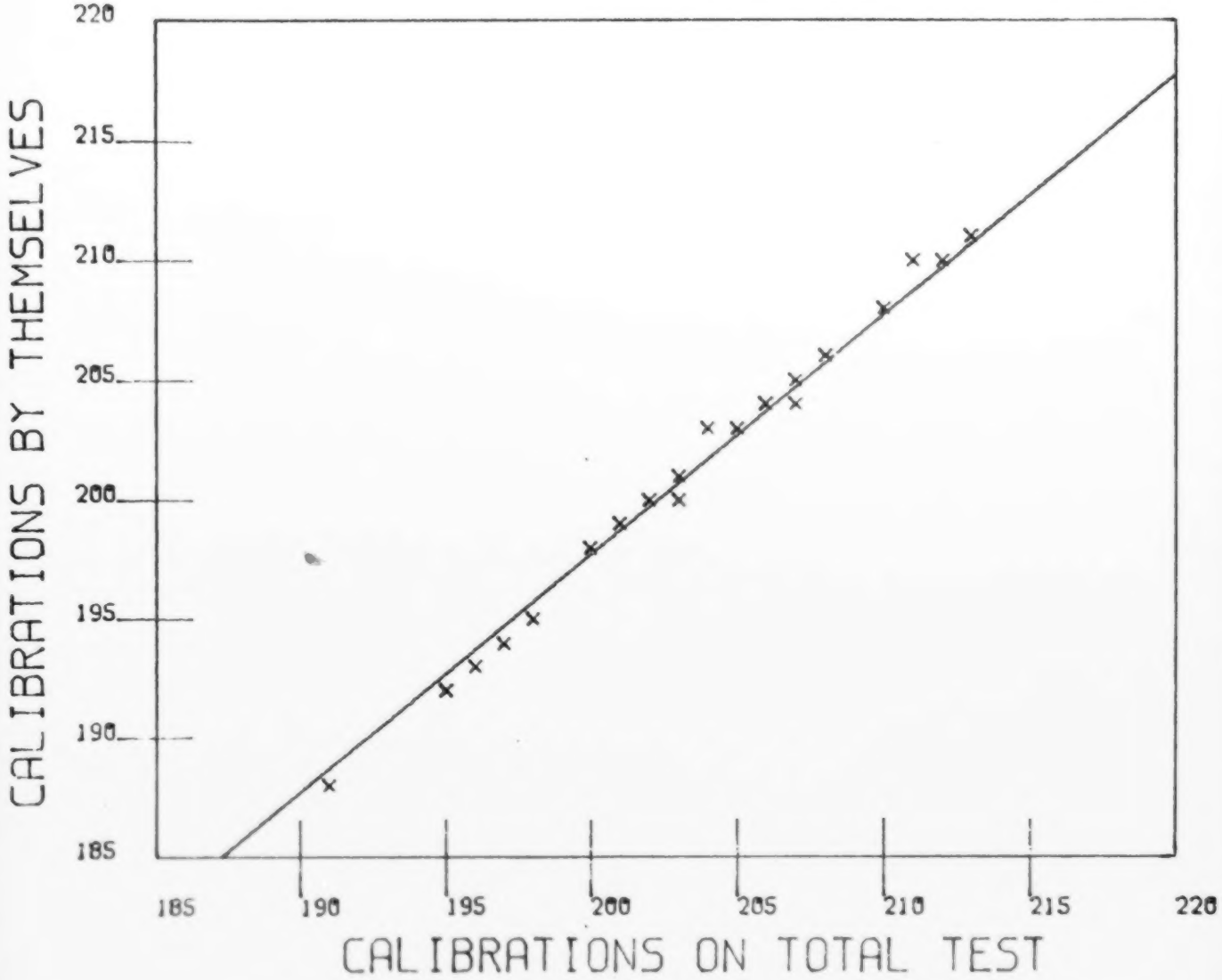

Figure 10. Effect of item discrimination. Items with high point biserials.

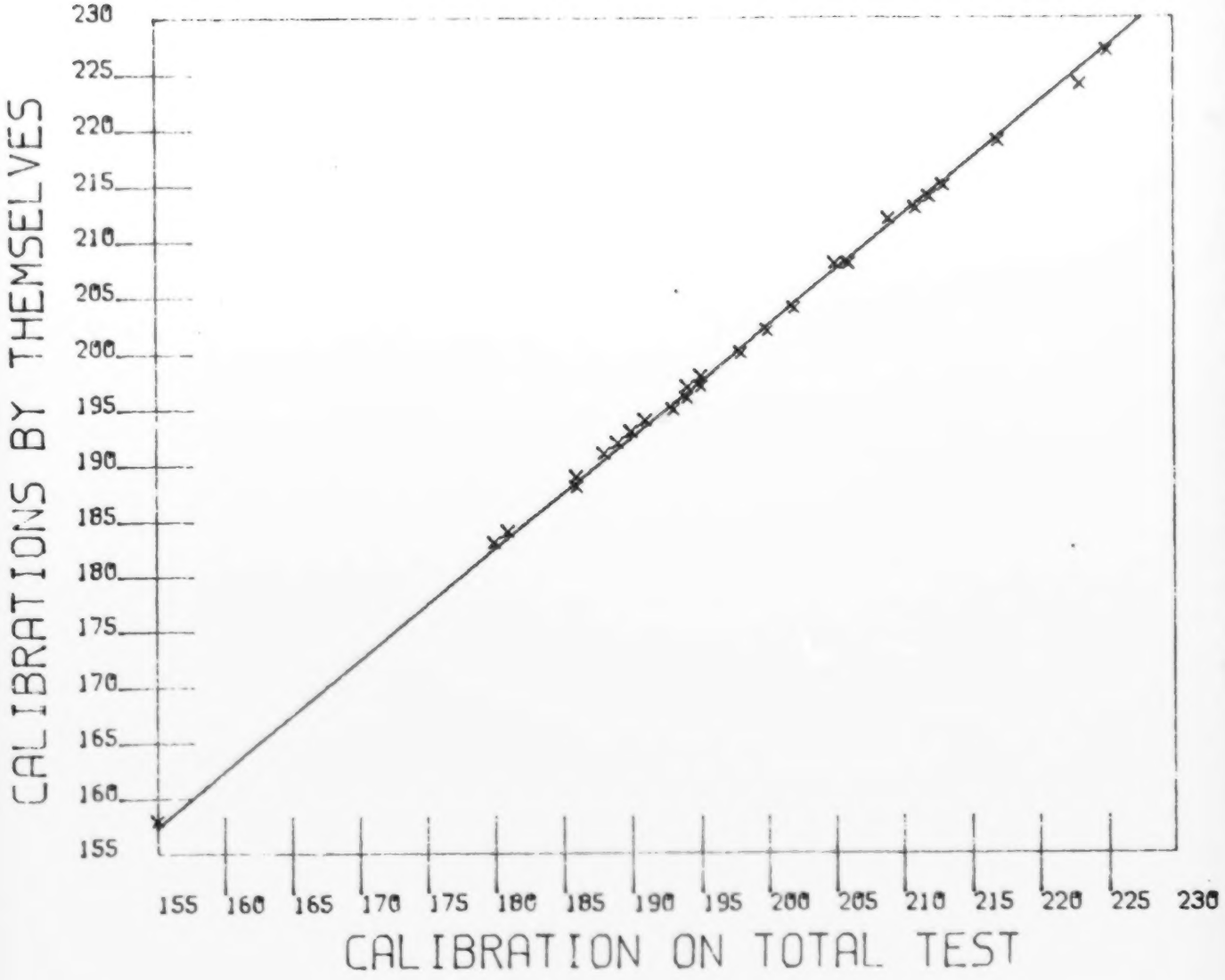

Figure 11. Effect of item discrimination. Items with low point biserials.

There is probably more to this story. These results of this study are based on multiple choice questions. Since these questions are screened before using them in a test, we have already eliminated most or all the items with low discrimination. Since point biserials for multiple choice items rarely exceed .5, there is a limit on the high end. What if we had examined open ended questions with extremely high point biserials? That question remains unanswered for now.

How Does Sample Size Affect Calibration?

This is an important practical question. When you build an item bank, you need lots of help from teachers and students. If scores from 200 students can give "accurate enough" calibrations, you move three times faster than if you need scores from 600. (This issue is discussed in Forster, 1975b; and Forster, 1978d.)

But what is "accurate enough"? We used the notion of diminishing returns. When the increase in accuracy does not justify the cost of an increase in sample size, then we have the "accurate enough" point.

The study we devised used 1,400 scores on a fourth grade mathematics test and 1,400 scores on an eighth grade reading test. We selected random samples of 50, 100, 200, and 300 scores. Then we calibrated the test for each sample (five samples of each size) and checked the difference between the calibration for the sample and for the total. The results of this procedure are shown in Table 10.

The breaking point seemed to be around 200 to 300. To pin this down further, we replicated the study with 3,800 fourth graders and 3,800 seventh graders on both reading and mathematics. The results of this analysis are shown in Table 11.

Reviewing these results, we concluded that a sample size of 250 to 300 strikes a good balance between accuracy and the practical requirements of field testing.

Does Grade Level Affect Calibration?

Sometimes grade level has an obvious effect on calibration. For example, some third grade students demonstrate high mathematical ability. It would be a mistake, however, to test them with questions on fractions, percents, and decimals when they haven't been taught those concepts. In fact, it is important to make sure students have been taught a concept before they are tested on it.

Can grade level affect calibration in more subtle ways? If so, what adjustments need to be made in building a bank?

The study we designed used 20,000 scores from the P.A.L.T. program. This testing program assigns the level of test in a subject based on each student's achievement. Therefore, any given test level is taken by students in several different grades with similar levels of achievement.

These tests were calibrated separately for each grade level. Then the calibrations were matched across grades. Finally, we calculated the absolute value of the deviations of each calibration for each grade from the average across grades. The results of this analysis are shown in Table 12 and are based on calibrations for groups of 600 or more students at each grade level.

Table 9

Effect of Restrictions in Item Level on Calibration

Middle Items#	Full Test	Middle 60	Middle 50	Middle 40	Middle 30	Middle 20
60	.998*	X	X	X	X	X
50	.996	.999+	X	X	X	X
40	.995	.999	.999+	X	X	X
30	.993	.999	.999+	.999+	X	X
20	.989	.997	.999	.999	.999+	X
10	.991	.998	.999	.999+	.999+	.999+

* Entries are correlations of calibrations.
# The smaller the number of items, the more restricted the range of calibration

Table 10

Effect of Sample Size on the Accuracy of Calibration

	4th Grade Mathematics		8th Grade Reading	
No. of Items	81		100	
Population	1478		1808	
Std. Deviation	14.15		15.25	
Sample Size*	Correlation	Average Absolute Val. Difference	Correlation	Average Absolute Val. Difference
50	.956	3.108	.965	3.042
100	.978	1.950	.982	2.074
200	.987	1.314	.989	1.381
300	.989	1.138	.991	1.103

* Entries based on five samples drawn of each size.

Table 11

Effect of Sample Size and the Accuracy of Calibration (Replicated)

	4th Grade Reading		4th Grade Mathematics		7th Grade Reading		7th Grade Mathematics	
No. of Items	35		37		36		30	
Population	3771		3827		3846		3773	
Std. Deviation	13.51		11.84		6.17		8.30	
Sample Size*	Corr.	Avg. Abs. Diff.#	Corr.	Avg. Abs. Diff.	Corr.	Avg. Abs. Diff.	Corr.	Avg. Abs. Diff.
50	.956	3.367	.950	3.291	.939	2.689	.901	2.531
100	.977	2.430	.975	2.201	.961	2.086	.941	1.884
150	.986	1.822	.985	1.577	.978	1.463	.955	1.639
200	.990	1.623	.990	1.388	.983	1.299	.971	1.244
250	.992	1.353	.992	1.173	.982	1.407	.972	1.219
300	.994	1.148	.994	1.113	.986	1.170	.978	1.067

* Entries based on five samples drawn of each size.
# Average absolute value differences.

As the averages for each grade and total show, there were two trends. First, as expected, the absolute value differences were largest for the extreme grades. Second, the differences were greatest for tests at the extreme, i.e., level 1 (lowest) or level 7 (highest). (This effect is obscured by the absence of some upper grades for the lower test levels and the absence of some lower grades at the upper test levels.)

The averages of the absolute differences were small and the differences were symmetrically distributed. So the impact of these differences on the scaling of the test would be minimal. But a few items demonstrated large differences in calibration across grades. For example, the following question had the largest average absolute value difference for reading:

Which statement is an opinion?

1. Newspapers publish advertisements.
2. Everyone should read the newspaper.
3. It takes many skills to print a newspaper.
4. Bi-weekly newspapers are printed twice a week.

This question had a calibration of 220 for grade three, 215 for grade five, and 211 for grades six to eight, indicating it presented special difficulty to the lower grade students even though their overall achievement was equivalent to the upper grade students.

Tracking down these differences systematically and relating them to the instructional program is beyond the scope of this paper. But this type of curriculum effort can be expected to yield important insight into student learning as well as testing.

Does grade level affect calibration? Yes, by increasing exposure to some instructional materials, and by influencing the way a student approaches some questions. Although this influence appears to be specific and limited, Coffman (1986) pointed out that monitoring these effects over time is an important professional responsibility of the test developer.

Does Time of Year Affect Calibration?

This question may seem like an instant replay of the calibration for different groups. There is no reason to expect calibrations to drift from fall to spring. But it would be reassuring to know for sure.

Ron Houser and Gage Kingsbury (1986) analyzed results for five years of spring and fall Portland Achievement Level Test scores for reading, mathematics, and language and concluded "...changes in item parameter estimates have not had any practical impact upon student scores over the period of the study."

We designed a study to compare calibrations for Level Tests taken in the fall in one district to the calibrations for the same tests taken in the spring in another district. The average absolute value differences and correlations of the calibrations fall to spring, fall to total, and spring to total are shown in Table 13.

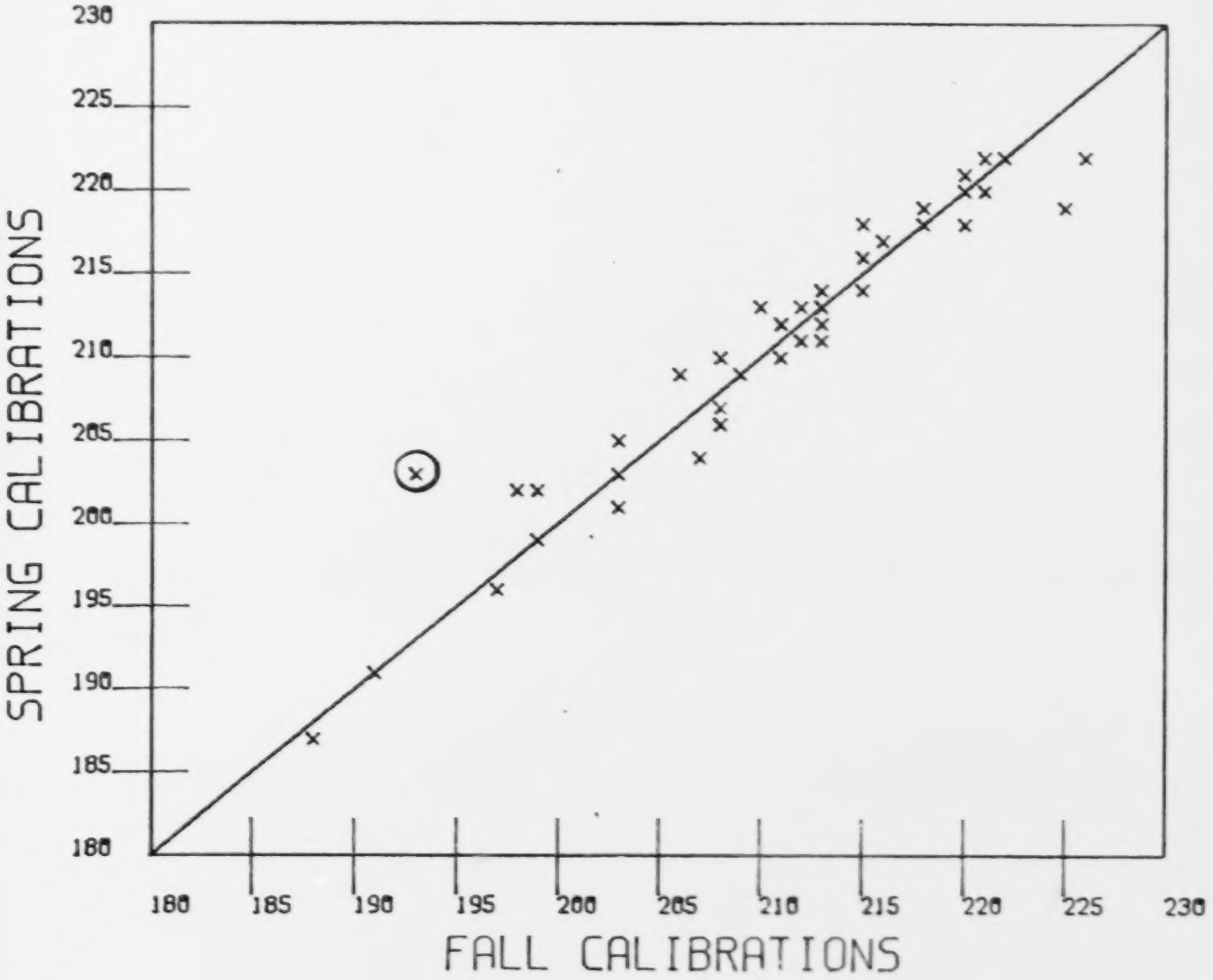

Figure 12. Fall vs. spring calibrations. Tests of Individual Performance Reading Level 5.

Table 13

The Effect of Time of Year of Testing on Calibration

Subject/Test Level	Average Absolute Value Difference	Correlation Fall/Spring	Correlation Fall /Total	Correlation Spring/Total
Reading/T02	2.43	.90	1.00	.94
/T03	1.55	.95	1.00	.97
/T04	1.97	.94	1.00	.96
/T06	1.38	.91	.99	.96
/T07	1.94	.92	.97	.98
Language/S23	2.56	.87	1.00	.90
/S24	2.02	.88	.99	.92
/S25	1.76	.89	.99	.94
/S26	1.90	.92	.98	.98
/S27	1.99	.91	.98	.97

Table 12

The Effect of Grade Level on Calibration for Different Test Levels on the Portland Achievement Level Tests

Subject/Test Level	Average Absolute Value Differences						
	Grade Level						
	Total	3	4	5	6	7	8
Reading/L1	1.09	1.05	.81	1.03	1.46	--	--
/L2	1.24	1.38	.71	1.95	1.07	1.07	--
/L3	1.43	2.05	--	2.07	.81	.86	1.36
/L4	.74	--	--	.81	.59	.64	.91
/L5	.86	--	1.27	.59	.59	.80	1.05
/L6	1.09	--	1.60	.87	.62	1.04	1.30
/L7	1.01	--	--	1.46	.68	.68	1.24
Reading/M2	.78	--	.83	.71	.80	--	--
/M3	1.03	--	1.46	.90	.73	.82	1.24
/M4	.81	--	--	1.19	.50	.67	.87
/M5	.76	--	--	1.17	.49	.62	.77
/M6	.78	--	--	1.16	.53	.53	.90
/M7	1.24	--	1.90	--	.78	.79	1.48
Language/T21	.75	.94	.54	.76	--	--	--
/T22	1.29	1.78	.76	--	1.08	1.52	--
/T23	1.56	2.10	1.01	--	--	1.29	1.84
/T24	1.46	1.90	1.04	--	--	1.29	1.61
/T25	1.02	--	1.49	.90	.50	.94	1.26
/T26	.77	--	--	--	.92	.49	.90
/T27	1.89	--	2.25	1.11	--	--	2.33
Language/U22	.91	--	1.03	.83	.76	1.00	--
/U23	.95	--	1.44	--	.57	.80	.99
/U24	1.28	--	1.82	--	--	.60	1.21
/U25	1.11	--	1.90	.79	.52	.98	1.36
/U26	1.72	--	1.75	1.26	--	--	2.15
/U27	1.72	--	1.84	1.11	--	--	2.22

As the correlations indicate, there was a high level of agreement between the fall and spring calibrations. Figure 12 shows the scatter-plot for fall to spring correlations for one of the tests. In this example, the "X" circled on the scatter-plot indicated that the following item had a spring calibration that was 10 Rasch units (Rits) higher than its fall calibration:

Which statement is a fact?

1. John is nice.

2. Steak is better than chicken.

3. Mary's hair is red.

4. Mary wears beautiful clothes.

As with grade level differences, the interpretation of this calibration shift has the potential of revealing important characteristics about the district instructional program. The results of this study, however, did not reflect a difference between fall and spring calibrations larger than that attributable to sampling error.

Summary

The research reported here on the Rasch model was designed with the demands of item banking in mind. Real tests were used as part of real test programs. The program designers were committed to in making the results useful to teachers. Teachers and students were conscientious in carrying out the testing. And everyone had a stake in the outcome. (A discussion of the procedures appears in Hathaway, et al., 1985.)

The research was carried out over twelve years. The researchers were not pressured to move on to new, hot, or sexy topics to increase their list of publications. But they knew that they had a professional responsibility to answer real questions about the reliability and validity of the tests they were developing.

Many hands shared in the effort. Fred Forster and George Ingebo of Portland Public Schools were chief architects for the research reported in this paper, with vital contributions from Victor Doherty, Dean Forbes, Walter Hathaway, and Ronald Houser. Beyond Portland, the staunch group of contributors included Bruce Arnekiev (Tacoma, Washington Public Schools), John Davis, James Moriel, and Paul Fracoli (San Jose, California Public Schools), Marshall Herron (Corvallis, Oregon Public Schools), Chad Karr (Portland State University), Clarence Mershon (Parkrose, Oregon Public Schools), Ray Miller (Northwest Evaluation Association), Wayne Neuburger (Oregon State Department of Education), Allan Olson (Hillsboro, Oregon Public Schools), and Peter Wolmut (Multnomah County Educational Service District). Ben Wright played a unique role bringing the Rasch model to our attention and then provided valuable consultation and inspiration. (Without Ben Wright, there wouldn't be a Rasch tiger to ride.) And there were hundreds of teachers and thousands of students in Portland and other districts from California to Washington to Michigan who provided the data that made the research possible. These efforts have directly benefited students, parents, teachers and administrators in the districts using the P.A.L.T. program.

The research and item bank building reported here occurred independently of the heated, and often unenlightened, debates, computer simulations, and data-free arguments of the academics which have usually proven counter-productive and misleading to our colleagues. The emotion demonstrated on both sides of these discussions often resembles theology more than science.

The Rasch model has proven to be a powerful development tool, when it is used properly. As Peter Wolmut (1974) pointed out, "The Rasch model allows us to have available new methods in solving our problems and these methods are superior in their refinement to what we have always had. With recognition of just what the Rasch model can do and with the collective inventiveness that is available, school districts will profit from superior test development based on Rasch calibrations."

Most arguments about the Rasch model are based on a double standard. We are supposed to worry about unidimensionality, independence of item order, and equal discrimination when applying the Rasch. Yet these same assumptions are implicit in classical test theory. And no one raises an eyebrow when publishers establish their norms on averages and standard deviations of raw scores which are ordinal and not equal interval. On the other hand, we are urged to embrace the two and three parameter models, even though there is no evidence that the parameters are sample independent and reproducible. Beyond this issue, imagine the practical problems posed by models which require us to explain to a parent why two children with the same number of correct answers receive different scores.

These arguments spring from a misconception of the potential of the Rasch model. Rather than a repudiation of classical theory, the Rasch model is its fulfillment. As experienced practitioners know, when classical analysis and Rasch analysis are prudently applied, the results agree. In fact, the Rasch model makes it possible to research questions and assumptions that we have had to accept on faith using the classical model. On the other hand, the Rasch model is conservative when contrasted with the radical, complicated, and narrowly researched two and three parameter models.

Item discrimination and test reliability were landmark inventions that revolutionized test analysis. Equal interval scores and item calibration have given the Rasch model a special place in history. To re-phrase Marshall McLuhan, those who focus on the medium will miss the message. The medium may be the Rasch model, but the message is school testing that is more accurate and useful to teachers, more humane to students, and more valid for program evaluation.

REFERENCES

Arneklev, B. (1979). Comparison of calibrations for different ethnic and language groups. Occasional paper, Tacoma Public Schools Evaluation Department, Tacoma, Washington.

Coffman, W. E. (1986). Taming the Rasch tiger: Discussion. In J. Hall & P. F. Wolmut (Eds.), National Association of Test Directors 1986 Symposia. Oklahoma City, OK: National Association of Test Directors.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. Journal of Educational Measurement, 23, 283-298.

Forbes, D. W., & Ingebo, G. S. (1975). An empirical test of the content homogeneity assumption involved in Rasch item calibration. Northwest Evaluation Association Journal, 1(2), 24-32.

Forbes, D. W. (1986). The Rasch model as a practical and effective procedure for educational measurement. In J. Hall & P. F. Wolmut (Eds.), National Association of Test Directors 1986 Symposia. Oklahoma City, OK: National Association of Test Directors.

Forster, F. (1975a). Using the Rasch model to improve school testing programs. Northwest Evaluation Association Journal, 1(1), 13-15.

Forster, F. (1975b). Optimum size for Rasch field testing. Northwest Evaluation Association Journal, 1(2), 15-18.

Forster, F. (1976). Can Rasch scaled scores be predicted from a calibrated item pool? Occasional Paper No. 7, Portland Public Schools, Portland, Oregon.

Forster, F. (1977). The application of the Rasch model. Northwest Evaluation Association Journal, 2(2), 33-40.

Forster, F. (1978a). Can Rasch item levels be determined without random sampling? Northwest Evaluation Association Rasch Monograph Series, 4-7.

Forster, F. (1978b). Is the Rasch item level scale equal interval? Northwest Evaluation Association Rasch Monograph Series, 8-12.

Forster, F. (1978c). Do students receive the same Rasch achievement level from different tests? Northwest Evaluation Association Rasch Monograph Series, 13-16.

Forster, F. (1978d). What is the smallest sample size needed for field testing? Northwest Evaluation Association Rasch Monograph Series, 17-20.

Forster, F. (1978e). Is the calibration of an item affected by the range of items on a test? Northwest Evaluation Association Rasch Monograph Series, 21-22.

Forster, F. (1978f). Mathematics of the Rasch model. Northwest Evaluation Association Rasch Monograph Series, 23-37.

Forster, F. (1982a). The fixed parameter model: a more accurate and flexible tool for item calibration. Occasional Paper No. 23, Portland Public Schools, Portland, Oregon.

Forster, F. (1982b, June). The fixed parameter model. Paper presented at the annual meeting of the Canadian Educational Research Association Conference, Nova Scotia.

Forster, F. (1986, June). The microcomputer can be used to improve testing--right now. Paper presented at the 7th National Educational Computing Conference, San Diego, California.

Forster, F., & Karr, C. (1986, May). Build your own calibrated item bank. Paper presented at the spring conference of the Washington Educational Research Association, Vancouver, B.C.

Hathaway, W. E., Houser, R. L., & Kingsbury, G. G. (1985). A regional and local item response theory based test item bank system. Occasional paper, Portland Public Schools, Portland, Oregon.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. Journal of Educational Measurement, 19, 139-147.

Houser, R. L., & Kingsbury, G. G. (1986, April). Assessing the stability of item response functions across time. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California.

Hunt, B. (1986). Implications of Rasch calibrated item banks for measurement of the locally planned curriculum. In J. Hall & P. F. Wolmut (Eds.), National Association of Test Directors 1986 Symposia. Oklahoma City, OK: National Association of Test Directors.

Ingebo, G. S. (1976). How to link tests to form an item pool. Occasional paper No. 6, Portland Public Schools, Portland, Oregon.

Ingebo, G. S. (1978). Historical perspective on the application of the Rasch model in the Portland metropolitan area. Northwest Evaluation Association Rasch Monograph Series, 1-3.

Ingebo, G. S. (1980). Some perspectives on equal interval curriculum scaling. Northwest Evaluation Association Journal, 5(1), 33-36.

Ingebo, G. S., Arneklev, B., & Gee, D. (1976). Optimum range of difficulty for linking items. Northwest Evaluation Association Journal, 2(1), 23-32.

Karr, C. (1976). Evaluation by triangulation. Occasional paper No. 4, Portland Public Schools, Portland, Oregon.

Mershon, C. (1986). How Parkrose improved its testing program using the Rasch model. In J. Hall & P. F. Wolmut (Eds.), National Association of Test Directors 1986 Symposia. Oklahoma City, OK: National Association of Test Directors.

Rentz, R. R., & Bashaw, W. L. (1977). The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 14, 161-179.

Slind, J. A., & Linn, R. L. (1977). Vertically equated tests: Fact or phantom? Journal of Educational Measurement, 14, 23-32.

Thomas, G. P. (1986). Information needs within a multi-district environment. In J. Hall & P. F. Wolmut (Eds.), National Association of Test Directors 1986 Symposia. Oklahoma City, OK: National Association of Test Directors.

Wolmut, P. F. (1974, May). Some uses and limitations of the Rasch model. Paper presented at the spring conference of the Washington Educational Research Association, Seattle, Washington.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In Proceedings of the 1967 computerized adaptive testing conference on testing problems. Princeton, N.J.: Educational Testing Service.

Wright, B. D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B. D. & Stone, M. H. (1979). Best test design : Rasch measurement. Chicago: MESA Press.